Operational Test and Evaluation of Artificial Intelligence-Enabled Capabilities

What to Consider in a Test & Evaluation Strategy

April 2024

Cleared for Public Release

This document provides a foundational overview how leveraging artificial intelligence (AI) in DoD capabilities will influence operational test and evaluation (OT&E) considerations in DoD T&E Strategies (TESs).

The OT&E of AI-enabled capabilities (AIECs) is necessary to build justified confidence that DoD systems will perform adequately when faced with the complexities and nuances of operational use.

Table of Contents

T&E Strategies for AIECs

This Section:

+ Specifies the role of the current document within the larger framework
+ Provides an overview of the framework for the test and evaluation of AI-enabled capabilities produced by CDAO Assessment and Assurance

01

This document is part of a framework for the T&E of AI-enabled capabilities

CDAO Assessment and Assurance is creating a framework to provide guidance on how to test and evaluate (T&E) AI-enabled capabilities (AIECs).

What is the framework?

The T&E of AIEC Framework provides best practices and guidance on how to test and evaluate AIEC.

The framework is organized into four categories of testing and provides different types of resources to AIEC developers and working-level testers.

Why is it needed?

The DoD community for the T&E of AIEC comes from a variety of backgrounds.

The T&E of AIEC Framework promotes a shared understanding between AIEC experts new to T&E and to T&E experts new to AIEC.

What is this document?

This document discusses what aspects of operational T&E (OT&E) to consider in a Test & Evaluation Strategy (TES) for an AIEC.

It is intended to help AIEC developers and working-level testers incorporate operational realism into testing throughout an AIEC's lifecycle.

This document provides:	This document does NOT provide:
✓ **Guidance and best practices**	✗ **Binding policy and requirements**
✓ **A primer on OT&E of AIECs**	✗ **A comprehensive OT&E guide**
✓ **Strategy-level T&E considerations**	✗ **Detailed T&E implementation**
✓ **T&E at the deployed, systems level**	✗ **T&E at the algorithm level**

CDAO's T&E of AIEC framework is organized into four focus areas

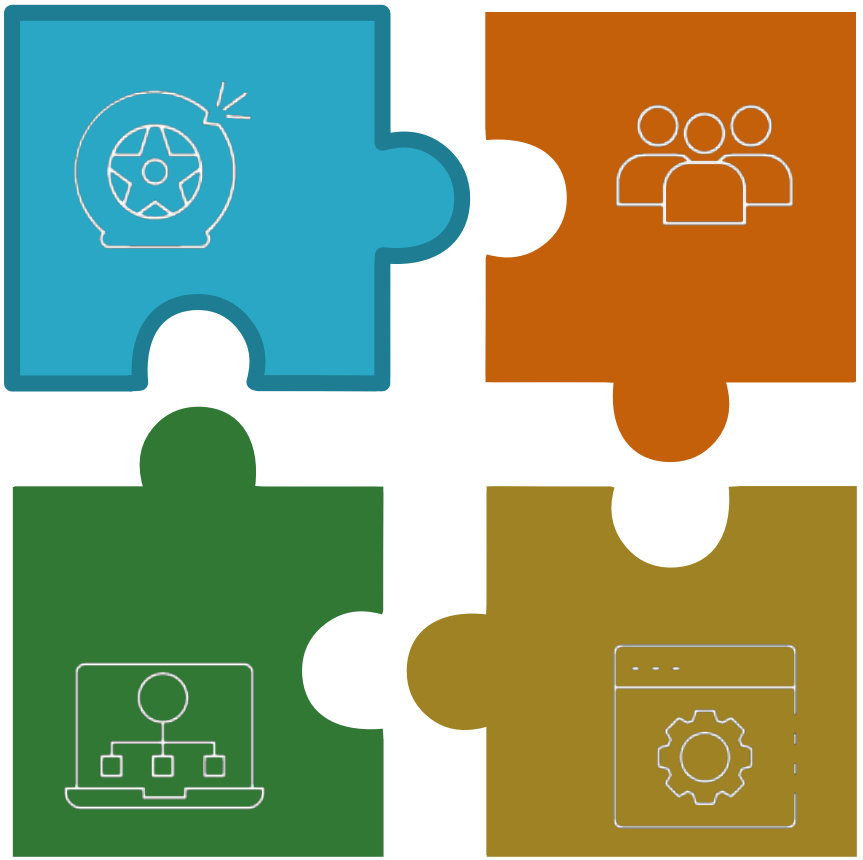
While these T&E focus areas help break critical aspects of T&E into digestible pieces, they are neither mutually exclusive nor cleanly delineated in real testing.

Operational T&E (OT&E)
Evaluating an AIEC performing representative missions within an operationally realistic environment against a realistic adversary.

Human Systems Integration (HSI) T&E
Evaluating an AIEC's ability to help stakeholders observe and orient to their environment, make informed decisions, and carry out their missions.

Systems Integration (SI) T&E
Evaluating an AI component within its larger system to ensure that the AIEC functions as a holistic unit and identify its limitations and risks.

AI Model T&E
Evaluating and documenting AI models and data across performance dimensions informed by system and mission constraints.



This document covers the OT&E focus area

CDAO is developing a series of products that address critical T&E needs

Part 1 is designed to help testers understand core T&E concepts so that working-level testers can write and assess test and evaluation strategies for AI-enabled capabilities

This document focuses on Part 1

1 | Write and assess T&E Strategies

Provides a high-level overview of critical T&E concepts that will be influenced by the inclusion of AI models in the system under test.

Supports testers and developers as they write TESs and assess whether the TES is committed to the right evaluations.

2 | Write and assess Detailed Test Plans

Provides guidance for implementation of T&E concepts introduced in Part 1; highlights promising paths forward for unsolved challenges.

Supports testers and developers as they develop and implement detailed test plans that capture mission objectives.

3 | Engage with other DoD T&E stakeholders

Provides frameworks outlining how T&E is critical to fielding trustworthy AIECs across DoD acquisition pathways and mission applications.

Supports testers and developers as they advocate for policy and investments that address DoD T&E shortcomings.

4 | Execute tests and rigorously analyze results

Provides resources such as templates, validated measurement instruments, and automated analysis tools.

Supports testers and developers by streamlining and automating common T&E activities with tailorable tools.

What is a Test & Evaluation Strategy?

A high-level document in DoD acquisitions that guides test planning and execution.

Captures the mission(s) a capability is intended to perform and all hardware and interfacing systems in the test design.

Identifies and prioritizes assessment areas to inform test team data requirements to support major program decisions.

Specifies the resources required to conduct T&E and shortfalls in resourcing that will require investments.

Describes the test events and activities necessary to evaluate the system and support acquisition, technical, and program decisions.

Learn More

You can read more about DoD TESs at https://www.test-evaluation.osd.mil/T-E-Enterprise-Guidebook/

OT&E is Important

This Section:

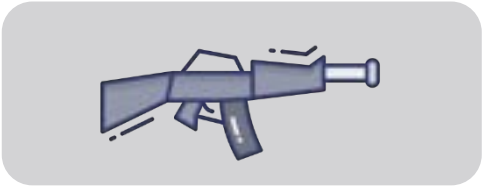
+ Introduces the unique test design considerations and evaluation *strategies for* OT&E
+ Illustrates the value of operational testing through historical examples
+ Provides a brief overview of OT&E oversight

02

Incorporating operational realism into testing saves lives, time, and money

History is full of examples where testing with representative environments and end users revealed—or could have revealed—system issues prior to fielding.

Operational performance consists of the interactions among the AI model, the units and end users employing it, and the mission objectives within a cyber-physical operational environment. You cannot assume performance observed on the AI model in insolation, nor the performance of the full AIEC in a controlled lab test, will be the same as in real operational use.

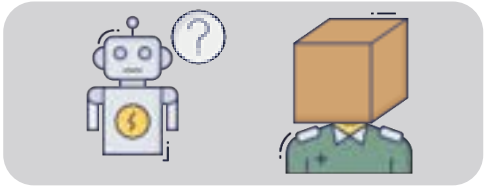


Failing to test in battlefield conditions can cost lives

The Vietnam War–era XM16E1 jammed frequently, endangering the infantry men it was designed to protect. A major cause of jamming was the gunpowder used in fielded ammunition, which was different from what the rifle was designed to use and was not required during acceptance tests. Non-representative testing obscured what could have been obvious beforehand: the XM16E1 was not suitable for the conditions of the Vietnam War.

Catching deficiencies in testing allows time for updates

During live fire T&E, Mine Resistant Ambush Protected (MRAP) vehicles were survivable against small to medium threats, which was in line with program requirements, but they had significant vulnerabilities against larger explosive threats that were more operationally realistic. To improve survivability, the MRAP Joint Program Office changed the vehicle's design prior to deployment.



Operational testing can reveal system limits

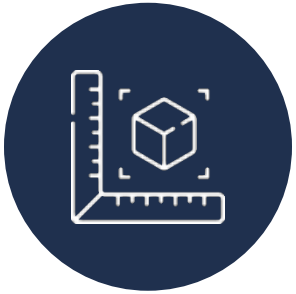
The DARPA Squad X program included AI designed to detect enemy forces in complex urban environments. The AI had been trained on images of soldiers walking and was deceived by soldiers moving in unconventional ways. Eight Marines were challenged to touch the AI sensor without being detected. All succeeded. They reached the sensor by somersaulting, hiding under a cardboard box, and camouflaging as a fir tree.

Incorporating operational realism into testing can be difficult

Many deficiencies and failures are often missed in earlier testing because the performance degradations stem from the application of the system, not an internal failure.

Operational Scope

Operational testers are often faced with complex operational spaces and sets of potential conditions for operational employment, including representative missions, users, workflows, interfacing systems, adversaries, and more.



Characterizing Systems

Beyond checking that a system meets requirements, OT&E often aims to characterize system performance across the operational space to understand how various conditions influence system performance.

Huge Testing Space

Attempting to characterize a system across the large scope of operational factors results in a very large testing space.

Realistically, testers cannot fully cover the testing space. They using rigorous test designs, relevant resources, and non-technical evidence to build a body of evidence.

OT&E must cover complex, expansive operational test spaces

OT&E generally involves more complex state spaces than other T&E, so doing it well requires intentional and nuanced test design and execution.

Test Design

Like all T&E, OT&E makes use of design of experiment (DOE) methods. However, the differences between OT&E and other T&E drive the challenges in applying DOE methods.

To determine which conditions are tested and how much testing is required, testers often employ DOE methods to maximize the utility of test data while defining a trade-space so that test resources are not wasted. Further, DOE helps testers balance factor levels and minimize correlations between test points.

DOE is not a one-size-fits-all tool, but a suite of tools practitioners pick from to fit their specific case.

Test Resourcing

Testing with representative missions, users, workflows, systems, and threats can increase the resources needed during testing and limit testing methods to those that minimize disruptions to realistic operational use during the test.

Realistically, testers do not have the resources to fully cover the operational space. Instead, they conduct testing on a subset of the operational space.

In some cases, methods that are “nice-to-haves” for traditional systems become necessary for AIECs (e.g., built-in telemetry to collect test data). Testers should begin planning early in the program’s lifecycle to take advantage of built-in infrastructure for recording data.

Non-Test Evidence

For T&E of AIECs, testers may need to understand and assess aspects of systems that have not traditionally been part of T&E, such as training data and descriptions of AI models. Accessing these resources may be particularly difficult for government testers, who have had issues when the government does not own technologies or allow for access during contracting.

Evaluations should not be limited to quantitative measurements; other factors may influence how to prioritize test activities. For instance, the quality of data provenance and curation impacts the risk of data poisoning.

Some DoD systems must undergo formal OT&E prior to fielding

This document focuses on the benefit of incorporating operational realism into testing across an AIEC's lifecycle; however, the next 2 pages discuss formal OT&E.

Some systems must be formally tested in an operational context prior to fielding.

All systems should incorporate operational realism into their testing across the system's lifecycle; however, some DoD systems are required to include formal OT&E in their test plans.

"The Director, Operational Test & Evaluation (DOT&E) … is the principal official and adviser to the Secretary of Defense on all DoD matters related to operational (OT&E) and live fire test and evaluation (LFT&E) of DoD systems and services acquired via the Defense Acquisition System" (DOT&E Website).

Programs on the Major Capability Acquisition (MCA) pathway and programs selected by DOT&E are subject to OT&E Oversight, as pursuant to sections 139, 4171, 4172, and 4231 of Title 10, U.S.C.

OT&E allows us to assess and anticipate how a fielded system will perform.

Primary Goal: Characterize a production-representative system to inform fielding decisions. OT&E reveals a system's capabilities and limitations to provide stakeholders with insight into how a system will perform under operational conditions. Testers measure various aspects of the system under test to characterize its performance and to evaluate its effectiveness, suitability, and survivability within an operational context.

Secondary Goal: Identify problems to address prior to fielding, such as by changing the system design or the tactics, techniques, and procedures. Systems with considerable problems often require additional OT&E after fixes are made.

Operational context includes a representative system, end users, and test environment.

"*The term 'Operational test and evaluation' means—(i) the field test, under realistic combat conditions, of any item of (or key component of) weapons, equipment, or munitions for use in combat by typical military users; and (ii) the evaluation of the results of such test*" 10 USC 139 - Director of Operational Test and Evaluation.

OT&E occurs within an operational context. All aspects of a test—from the mission, the end users, the integrated systems, and the threats—must realistically represent fielded operations for the system under test. The system under test must be production representative and accurately represent the planned fielded configuration that end users will use.

Formal OT&E includes four system performance attributes

Formal OT&E characterizes a system across multiple system attributes.

OT&E characterizes systems in context with users in context across these four attributes:

1. Operational effectiveness,
2. Operational suitability,
3. Survivability, and
4. Lethality.

Depending on the Service acquiring the system, the system under test, the test organization, and other considerations, OT&E may include each of these four attributes or a subset of these four attributes. At a minimum, OT&E evaluates operational effectiveness and suitability.

Successful OT&E does not mean that a system "passes" the test. Rather, a successful test is one that maximizes understanding of the system under test for each relevant system attribute.

Operational Effectiveness

The degree to which the system supports accomplishing mission objectives when used by representative personnel in a representative environment. Effectiveness measures are highly dependent on the system's intended use.

Operational Suitability

The degree to which the system can be satisfactorily placed and employed in the field. Suitability issues include reliability, availability, maintainability, usability, training requirements, transportability, safety, and other key elements.

Survivability

The degree to which the system and its personnel avoid or withstand a hostile environment that is operationally representative. Kinetic, non-kinetic, and cyber resilience encompass susceptibility, vulnerability, and recoverability.

Lethality

The degree to which a production-representative munition is effective against a threat-representative target within an operational environment. It is assessed in terms of hit probability and hit distribution.

Challenges of OT&E of AIECs

This Section:

+ Provides an overview of key AIEC characteristics and how these characteristics relate to operational testing
+ Discusses *long-standing OT challenges that have additional considerations for AIECs*

03

Operational realism is important for the T&E of AIECs

Most challenges associated with OT&E of AIECs are not unique to AIECs. However, the nature of AI can change and complicate problems that testers already face.

The goals of OT&E will not change for AIEC, but the technical and statistical approaches, test planning and assurance methods, design techniques, tools, and inferences that can be made must be updated in light of the way AIECs function.

AIEC can perform unpredictably when deployed outside of the conditions in which they were trained because of emergent model behaviors, and they can fail to generalize in ways that conform to human expectations.

It can be difficult, if not impossible, to evaluate which operational conditions are likely to cause system failures or undesired behavior. This reality makes it particularly challenging to determine an AI/ML model's robustness—that is, how effectively it performs in the face of noise—either inherent to the environment or generated by an adversary in an attempt to degrade performance.

A key concept for testers to consider is the difference between the performance of a component model and that of the overall system under test.

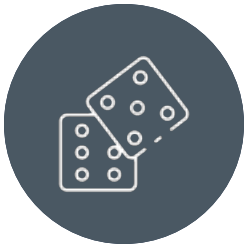
This framework identifies 5 OT&E challenges that are exacerbated by AIECs:

- Resourcing test activities,
- Generalizing from test to field,
- Characterizing causal relationships,
- Detecting and mitigating novel threats, and
- Tracking performance drift.

Testers of AIECs must be aware of these added challenges and plan accordingly. The following slides provide a quick overview of key concepts and considerations for each of the 5 OT&E challenges.

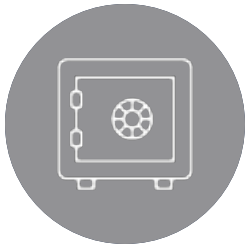
AIEC characteristics can exacerbate preexisting OT&E challenges

Testers must account for properties that are common to AIECs, such as complex, probabilistic decision-making and reward hacking.



Complex Decision-Making

Many AIECs have non-linear, non-continuous, and/or non-deterministic responses to stimuli that make system performance difficult to predict.



Black Box Algorithms

“Black box” models lack transparency, complicate the interpretation of decision processes, make defect diagnosis difficult, and hinder trust calibration.

Gamification & Reward Hacking

AI models are trained for specific performance measures. Misalignment of those measures with mission goals can lead to undesired behaviors and poor mission effectiveness.

Agile, Iterative Development

Like conventional software, AIECs may evolve and field updates over time. Pursuing iterative development requires continual tester engagement.

Overfit to Training Data

AIECs are often overly optimized to their training data. Brittle and hyper-tuned to training data, overfit models are ineffective in a real environment.

Learn More

See the **CDAO A2’s National AI T&E Infrastructure Capability (NAITIC) Gap Study** for further detail on how the inclusion of AI in DoD systems affects testing needs.

How to use this section

Each OT&E challenge is presented in a “one-pager.”

This framework identifies 5 different OT&E challenges that should be included in a TES.

Use this section to write or review a TES so that it includes core OT&E concepts relevant to AIEC.

How should I use this section?

Identify core concepts: In this product, we identify the critical OT&E concepts to consider when testing and evaluating AIEC.

Find “Google-able” terms: For each concept, the one-pager includes its more formal name and definition. Beyond being informative, this provides the keywords needed to find the supplemental literature online.

Learn to interpret informal language: Because many TESs will not have input from AI experts, one-pagers provide overviews and AI-specific concerns so that testers can identify if the TES has included relevant OT&E concepts with different, informal language.

Understand the need to test: We explain how each OT&E challenges can impact our ability to effectively, safely, and ethically employment these novel systems.

What are the limitations of this section?

It is not an exhaustive product.

While the core OT&E challenges included in this product highlight key issues that testers should focus on, please be aware that this list is not complete. While more nuanced concepts and implementation guidance will be discussed in future “guidebook” and “deep dives”, no product in this series will exhaustively list all OT&E concerns.

Additionally, these summaries are limited to a single page, but in reality, most of these concepts span entire research communities.

Not all AIECs will be impacted equally by these OT&E challenges.

Every TES may not emphasize the 5 OT&E challenges in this framework equally. Some will have to prioritize resources, and some challenges may be less relevant for some systems.

Resourcing test activities

Targeted factor selection and efficient experimental design are both increasingly important and difficult when systems include AI components.

How is it relevant to testing?

Operational testing is constrained by limited resources, such as time, money, expertise, and infrastructure. As a result, it is not feasible for OT&E to completely cover the operational envelope. Instead, robust DOE is crucial to maximizing the success of OT&E. Additionally, integrating DT and OT&E where appropriate could conserve overall testing resources.

OT&E of AIECs will place additional demands on personnel, funding, infrastructure, and time.

How can AI make it harder?

Compared to traditional systems, AIECs will demand more testing resources; frequent changes require frequent test events, and the operational envelopes are larger with unpredictable performance. This requires special tools and practitioners that understand how to measure AIEC behavior. Unfortunately, lack of expertise on AIECs limits our ability to meet such demands. AIEC characteristics make it difficult to determine how or whether the state space is adequately covered during OT&E. Because of black box algorithms and probabilistic decision-making, it can be unclear which factors will affect AIEC performance. Inconsistent performance over the operational envelope may require less conventional DOE, such as prioritizing covering edge cases and low-density training data regions.

What should testers do?

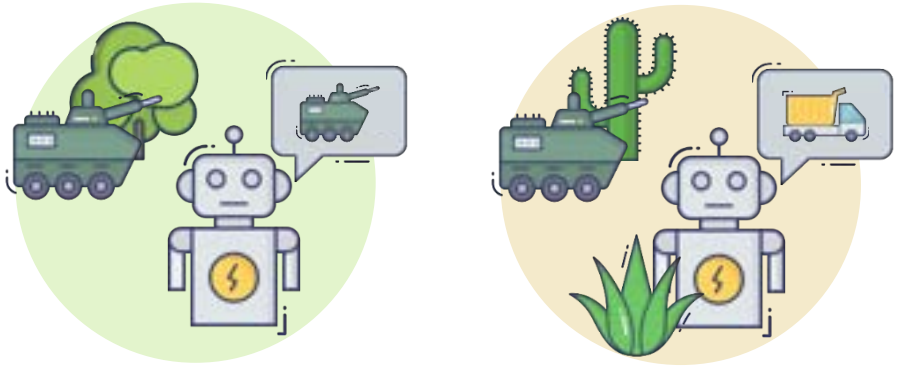
As with traditional systems, DOE and integrated testing will continue to be important for OT&E of AIEC even though they may look different. For example, training data and other developmental artifacts may inform OT&E factor selection. Testers will also need to develop new methods and metrics. Productivity boosting, such as through automated testing, could mitigate limited resources.

Generalizing from test to field

We use OT&E to characterize what fielded performance will look like; AIEC increases the difficulty of ensuring our characterizations are accurate.

How is it relevant to testing?

We make inferences about real-world operations by generalizing the results observed during operational testing. While the test results rarely match fielded operations exactly, we take steps to increase confidence in our inferences by testing representative systems within an operationally realistic context and by employing rigorous scientific practices like DOE.



Observed performance in one context might not translate to other environments or applications.

How can AI make it harder?

AIECs are often selected over traditional technologies for their ability to rapidly identify trends across vast quantities of data that would be overwhelming or nonintuitive to humans. However, an AIEC's ability to pick up subtle trends can also make it challenging to identify appropriate test factors and edge cases. Harsh environments and the complexity of many DoD applications can make it difficult to meaningfully evaluate an AIEC within a mission context. Furthermore, the T&E community has a less mature understanding of data features that influence an AIEC's performance for applications not common to industry. If an AIEC is trained with poor quality, non-representative data it will likely be ineffective; however, it may appear misleadingly effective if the test dataset is not sufficiently operationally realistic.

What should testers do?

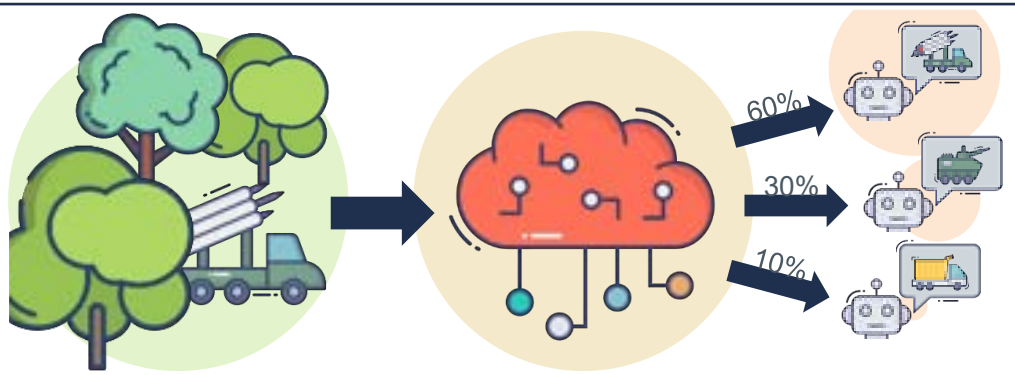
To make accurate generalizations, additional considerations, methods, and resources are required for AIEC as compared to traditional systems. Operational realism, such as in model training data, should be introduced early in the development cycle. Testers will need to account for additional uncertainty and will need to find new ways to make ranges operationally realistic for AIECs.

Characterizing causal relationships

Non-deterministic systems make it especially difficult to identify causal links between input data and system performance.

How is it relevant to testing?

Knowing causal relationships between inputs to the system and performance outputs helps testers predict system performance from test results. As systems become more complex, so too do causal relationships between input and output. This complexity challenges testers to identify factors most likely to affect system behavior, and how to use behavior observed in testing to predict behavior in untested scenarios.



With probabilistic decision-making, one input can result in multiple outputs.

How can AI make it harder?

AIECs that use probabilistic decision-making are non-deterministic in that a single input could result in multiple system performance outputs; this multiplicity makes characterizing causal relationships difficult compared to deterministic systems. While this problem is not unique to AIECs, it is more prevalent for AIECs and further exacerbated by black box algorithms. To predict AIEC performance, testing must collect a distribution of causal relationships that over time yields a characterization of system behavior uncertainty.

What should testers do?

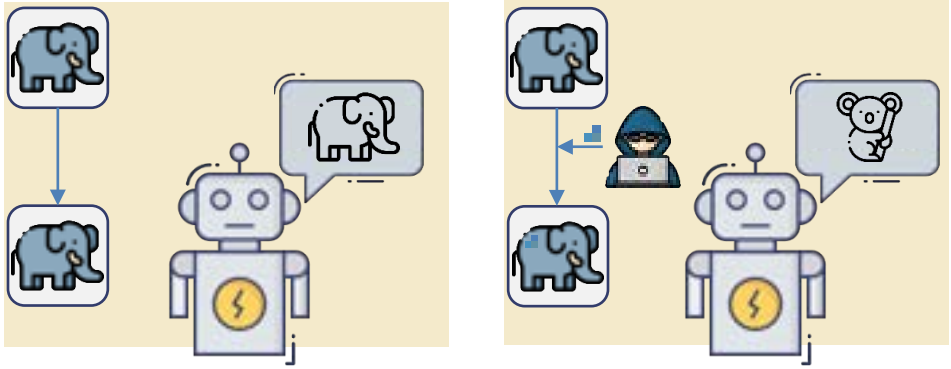
We need new T&E methodologies for non-deterministic AIECs. Standard T&E methodologies are insufficient because current testing assumes system performance can be definitively predicted based on outcomes from a small number of test points. Depending on the AIEC, assurance cases may be used to scope a set of possible failure modes.

Detecting & mitigating novel threats

Adversaries may utilize new attack vectors to compromise and exfiltrate data from critical AIEC.

How is it relevant to testing?

New technologies present new attack vectors. To accurately characterize a system's survivability, testers must keep up with and simulate ever-evolving ballistic, electronic, and cyber threats. Adversaries continually develop new methods and techniques, making it difficult to anticipate and test against all possible scenarios.



Adversaries can change subtle features that humans miss but that impact AI performance.

How can AI make it harder?

Because they are software intensive, AIECs have large cyber-attack surfaces. Some cyber-attack vectors are specific to AIECs. AIEC-specific vulnerabilities include extraction (where hackers query an AIEC to reverse-engineer sensitive information like training data) and data poisoning (where adversaries tamper with the datasets used to train and deploy an AIEC).

Both black box algorithms and probabilistic decision-making can make it difficult to anticipate how a system may be vulnerable. Additionally, with increases in autonomy comes an increased risk that human operators might not catch unplanned AIEC behavior. This risk gives more reason to set up a robust system for runtime monitoring and securing AIECs against tampering.

What should testers do?

Testing novel AIEC attack vectors will require cyber assessments of all networks and systems on which an AIEC operates. Testers should create adversarial examples and other robustness training techniques for the test and development of AIECs. Testers should develop metrics for assessing whether AIEC features are likely to decrease vulnerabilities.

Tracking performance drift

Over time, fielded performance may differ from observed test performance; this drift becomes increasingly likely with AIEC systems.

How is it relevant to testing?

Fielded performance may change over time for reasons including decreases in reliability, changes to real-world operations, and off-label use. Failure to monitor performance drift after fielding risks that our understanding of systems become increasingly unrepresentative over time. Testing post fielding can mitigate this risk, but it is not standard procedure for continually changing systems.

The environment, model, or application may change over time and result in performance drift.

How can AI make it harder?

Many factors contribute to performance drift of fielded AIEC systems. Changes in operational context may result in data drift, which can reduce model performance. Overfit and brittle systems exhibit poor performance in operational contexts with inputs that differ from the training data. Performance may drift due to unexpected results in the system's reward function, especially as the AI model and AIEC operators learn more about the system's internal reward structures. As with many software-enabled systems, AIECs will continue to update after fielding, but these updates may be more frequent and less noticeable than with conventional software. AIEC iterations may be both a response to and a cause of performance drift.

What should testers do?

To assess and account for performance drift, we should move toward continually measuring system performance. Testing post fielding may capture drift, though testers will need to consider when and how much to test fielded systems. Alternatively or in addition to testing events post fielding, testing may be automatically run with updates while using fewer resources than with repeated manual testing.

OT&E over the AIEC Lifecycle

This section:

+ Argues for AIEC OT&E to “shift left” by gathering data collected under operationally representative conditions in DT
+ Argues for AIEC OT&E to “shift right” by evaluating data from fielded systems and conducting follow-on testing to account for emerging behaviors

04

We must continue to "Shift Left"

An ounce of prevention is worth a pound of cure.

"Tech debt"—or the potential future costs incurred from shortsighted decisions made during development, which may need to be corrected or revised later—is a common challenge in technology development.

Development of AIECs can further intensify tech debt, particularly because of the close relationship between data quality and model performance.

Shifting operationally realistic testing earlier in the lifecycle can help reduce tech debt. Correcting a complex model trained on unrepresentative data may be exceedingly challenging, if not impossible, late in development.

"Shifting left"—or adopting an iterative, agile method to incorporate operational realism early and often—can help reduce the risks of building up excessive tech debt and training on unrepresentative data.

Shift Left

Operational realism must be incorporated into developmental testing

Operational environments and uses must be considered in developmental testing.

Examining AIEC performance earlier within a realistic operational context enables a better understanding of AIEC performance and decreases unexpected performance deviations compared to testing in a sterile, scripted developmental environment.

Furthermore, incorporating aspects of OT&E early in development supports building a body of evidence that supports fielding the system. Operationally realistic developmental testing will not completely replace rigorous end-to-end, mission-based testing for systems that require formal OT&E, but collecting operationally relevant data to learn about system progress toward effectiveness and suitability at earlier stages in development will help mitigate inefficiencies and inadequacies in testing.

Decisions made in development need to be documented so that tests can be scoped properly.

In an agile, iterative model development approach, AIECs are frequently updated. Developers need to track these changes and assess their impact on performance.

Evaluating the impact of an AI model on the performance of the system in which it is embedded can pose several challenges. Additionally, while AI models are often touted as being “data-driven,” the reality is more nuanced. Given resource constraints and the complexity of operational environments, developers of AI models make assumptions and tradeoffs based on characteristics of the input data and model selection.

Shift Left

Early knowledge of an AIEC's data and data pipelines should inform testing

Testers need an understanding of the training data to inform test design

One of the core unique aspects of AIECs is that the model is not explicitly programmed but rather is trained on a dataset. This dependence on a training dataset brings the need to ensure the training and test data are representative of operational conditions.

Models can be overly sensitive to small changes in inputs, meaning that test results in one environment may not be representative of performance in another environment. Characterizing the data delivered by operational data pipelines will improve the chance the algorithm performance is representative of actual operations. Characterizing these data needs to be done early enough to support timely OT&E of AIECs.

Robust data and data pipeline documentation are critical to assuring data quality

Datasets often are not static; tracking data evolution is crucial to ensuring that the model remains valid and accurate. Beyond quantifiable assessments of data content and quality, documentation of the data—including their provenance, their processing, and the boundaries of the environment—is needed to provide context, enable reproducibility of results, and maintain a chain of custody to ensure data integrity and security.

Documentation will need to be tailored to various stakeholders to enable developers to iteratively address gaps in performance and to allow end users to make informed decisions on when and how to use their AIECs.

Shift Left

Simulations and data generation are useful tools, but they are not a panacea

Modeling and simulation (M&S) can help move operational realism earlier.

AIECs can have such large operational spaces that they cannot be fully observed via testing. Underpowered sampling of this large and complex performance space is particularly a concern during live testing. M&S and virtual environments can augment the evaluation of system performance, improve the exploration of edge and corner cases, and prioritize the selection of live test points.

Given that AI models are prone to being overfit to training data, any M&S used for testing will need to be sufficiently realistic to accurately represent system performance. While the required level of fidelity will vary depending on a variety of factors, applications of M&S for the OT&E of AIECs generally are immature and require additional research.

Beyond testing, automated governance mechanisms may help buy down risk.

Automated governance mechanisms that prevent an AIEC from operating outside defined boundaries of acceptable behavior are commonly suggested to mitigate risks. For example, an autonomous car navigation system might have constraints to prevent it from exceeding speed limits or crossing double lines.

However, a poorly implemented governance mechanism might lead to undesirable outcomes, particularly in off-nominal situations, such as failing to cross a double line to avoid an accident or driving at the speed limit in snow. While automated governance mechanisms potentially may limit unsafe behaviors, their performance must be assessed to minimize unanticipated negative outcomes.

We must continue to "Shift Right"

Determining what constitutes adequate performance for tasks without historical benchmarks makes evaluating an AEIC's performance challenging. Testers need metrics and threshold criteria that are operationally relevant, traceable, and understandable to assess performance. Fielding an AIEC in phases may help buy down risk.

AIECs may require additional testing after fielding due to changes that cause performance drift, such as changes to the algorithm or data on which it is trained. New sensors, environmental changes, and model updates, may also impact performance.

Continuous monitoring is essential to detect performance degradations in the field. Testers can use data collected via monitoring and feedback from fielded systems to support ongoing, independent assessments. Periodic assessment by operational test teams of the fielded baseline provides an objective determination of capability improvement and continued security.

Testers should consider what outcomes warrant what types of interventions. Interventions should include criteria for recertification, such as milestone achievements or risk assessments.

T&E cannot stop at deployment.

We need a post-fielding TES.

Shift Right

It can be hard to know when an AIEC is mature enough to be fielded

It is hard to know when an AIEC has been adequately tested.

Traditional T&E methods are still relevant, but the non-linear, non-continuous, and/or non-deterministic nature of many AIECs—combined with their brittle nature—can make it difficult to determine when you have adequately tested the system.

Furthermore, AIECs are often tasked with work traditionally performed by humans, for which there may not be established evaluation standards. Determining what constitutes adequate performance for tasks without historical benchmarks, especially for tasks that are not easily quantifiable, makes evaluating an AIEC's performance challenging. Testers need metrics and threshold criteria that are understandable, operationally relevant, and traceable to assess performance.

Fielding an AIEC in phases may help buy down risk.

To progressively field a capability while minimizing risk, the following approaches can be applied:

Limited Capability Rollout: Deploy the system with certain features enabled while others remain locked or restricted. This controlled functionality allows for early detection and mitigation of potential issues without exposing the full capabilities to end users.

Selective User Testing: Deploy the system to a select group of users for beta testing. This group will use the system in real-world conditions, providing feedback and data that can be used to improve the system.

Important!

Identify & document "Shift Right" activities early and often!

Shift Right

AIECs can warrant T&E during fielding due to updates and unforeseen changes.

Changes to the AIEC or its environment after fielding may necessitate additional T&E.

Changes to the system that cause performance drift include the following:

- Changes to the AI model and
- Changes to the data on which the AI model is trained.

Changes may also be something other than AI model updates, such as:

- New or updated sensors that collect data for AI processing,
- Changes in the operational environment, and
- New or updated concepts of operations.

Unforeseen changes in the environment may impact AIEC performance.

An AIEC's performance can be influenced by unanticipated changes in its environment, such as:

- A dirty camera lens disrupts the AIEC's ability since it is not trained for variations in dirt accumulation,
- Upgrading to a lighter hardware component degrades an AIEC's performance,
- Seasonal changes in vegetation cause an AIEC to confuse dirt roads with fields in autumn, and
- An AIEC fails to classify adversaries correctly when they change their clothing.

Important!

Identify & document "Shift Right" activities early and often!

Ongoing monitoring is needed to detect performance degradations in the field

Learning new lessons will accompany the fielding of new AIECs.

As AIEC become operational, continuous system monitoring and user feedback are required to understand and promptly address emerging issues. To manage risk efficiently, testers should use data collected from fielded systems for ongoing, independent assessments.

Effective monitoring and data collection should be planned and integrated into the overall test strategy. Explicitly outlining how monitoring data supports OT&E ensures robust AI systems even after deployment. Collaboration between testers and program managers is essential, given the evolving landscape.

Some performance degradations may be severe enough to require intervention.

Performance degradation warrants intervention when it significantly impacts capability or security. Operational test teams periodically assess the fielded baseline to objectively evaluate improvement and security. These assessments guide the scope of future independent testing.

Periodic evaluations provide insights into capability improvement and continued security. If performance drift significantly impacts functionality or safety, it signals the need for corrective action. These assessments inform decisions on whether to roll back to a previous version, restrict functionality, or temporarily remove the system from operation. Clear criteria, user feedback cycles, and risk assessments guide these judgments.

Important!

Identify & document “Shift Right” activities early and often!

Shift Right

Performance degradations or anomalies in the field may require intervention

Interventions for poorly performing AIECs should be risk informed.

After the system has been fielded, changes may warrant recertifying the system. Some of these changes will be planned and others will result from monitoring and discovery after fielding.

Testers need to consider what outcomes warrant what types of interventions, such as when a fielded AIEC should be "rolled back" to a previous version, restricted to a subset of functionality, or temporarily removed from the field.

These actions should include criteria for recertification, such as milestone achievements (e.g., resolution of identified issues), feedback cycles with scheduled time for user feedback, and risk assessments to determine when the AIEC is ready to be fielded (again).

DoD acquisition pathways provide little to no guidance for continued T&E of fielded AIECs.

Until more guidance is published, testers and program managers will need to collaborate to answer the following:

- Is there a process for rolling back to a previous version if needed?
- Who is responsible for rollback processes?
- Who has authority to make these decisions?
- Do all of the individuals responsible for the processes described either have the required expertise or have access to the required expertise?
- Is there an independent audit of these processes?

Important!

Identify & document "Shift Right" activities early and often!

Shift Right

Reflecting on OT&E of AIECs

+ Discusses how successful OT&E is critical for deploying trustworthy AIECs

05

Framework Recommendations

1

OT&E saves lives, time, and money by revealing issues with a system prior to its fielding.

Prior to fielding a system, it should undergo T&E within an operational context—that is, a representative system, with representative people, and in a representative environment—to anticipate and assess how the fielded AIEC will perform.

2

TESs must adapt OT&E activities to account for novel challenges posed by DoD AIECs

Testing AIECs can exacerbate pre-existing OT&E challenges by:

- Demanding more resources when the current ones are already limited,
- Increasing the difficulty of accurately generalizing T&E performance,
- Obscuring causal links between input data and performance, and
- Requiring an understanding of novel attack vectors.

3

OT&E must be incorporated across the AIEC lifecycle, from acquisition to sustainment.

Shift Left—Bring more aspects of operational testing into earlier developmental tests.

Shift Right—Use appropriate follow-on OT&E plus on going evaluation with data from the field to mitigate the risks of undesired or degraded system performance.